FRAMEWORK
PARA DATA MINING EDUCATIVO:
FORMALIZACION Y APLICACIONES

Marcelo Omar Sosa[1], Carlos Iván Chesñevar[2], Eugenia Cecilia Sosa Bruchmann[1]

1Departamento Computación/Facultad de Ciencias Exactas y Naturales/Universidad Nacional de Catamarca
Av. Belgrano Nº 300 - Planta alta - C.P: 4700 - San Fernando del Valle de Catamarca
Teléfono: 0383- 4425610 /4420900

2Instituto de Ciencias e Ingeniería de la Computación (Conicet –U.N.S.) / Facultad de Ciencias e Ingeniería
Universidad Nacional del Sur
Av. Alem 1253 -B8000CPB Bahía Blanca, Argentina
Teléfono: 0291- 4595135 ext. 2610

sosamod1@hotmail.com, cic@cs.uns.edu.ar, sosab_ec@hotmail.com

Resumen

Las técnicas de data mining permiten analizar grandes volúmenes de datos en búsqueda de información oculta y relevante para la toma de decisiones. Estas se aplican en diversos campos en donde se almacenan grandes volúmenes de datos de las actividades realizadas y cuyo procesamiento no puede realizarse utilizando otras técnicas. En el caso de datos obtenidos de procesos educativos, éstos presentan características particulares que requieren técnicas y formas de interpretación de resultados especiales por lo que dio origen a la rama de data mining denominada Educational data mining o E.D.M. por sus siglas en inglés. El proceso educativo moderno incorpora la tecnología como medio de comunicación y de desarrollo de actividades fuera del ámbito del aula, si bien existen diferentes tipos, el más utilizado es el denominado blended learning ya que representa una adecuada combinación de actividades virtuales y presenciales con el objetivo de enriquecer el proceso. La actividad educativa así desarrollada genera grandes volúmenes de datos, su procesamiento con técnicas de data mining y la interpretación de los resultados obtenidos requiere la creación de un framework que agrupe las técnicas, prácticas y criterios que sean más adecuadas para el procesamiento de este tipo especial de datos y que ayuden al docente a su aplicación e interpretación.

Palabras clave: Técnicas de data mining, Educational data mining.

Contexto

El presente trabajo de investigación se desarrolla dentro de las líneas prioritarias del centro de investigación en estadística aplicada (CEA-Fa.C.E.N.-U.N.Ca.) que fuera creado en el año 2015 en la Universidad Nacional de Catamarca Facultad de Ciencias Exactas y Naturales en el marco del Sistema de Investigación, desarrollo e innovación. Además los docentes están a cargo de asignaturas de las carreras del área informática de la Facultad de Ciencias Exactas y Naturales *(Fa.C.E.N.)* de la Universidad Nacional de Catamarca *(U.N.Ca.)*. Estas asignaturas se dictan en el campus central en la Capital de la provincia de Catamarca como así también en las subsedes del interior (Departamento Ancasti) y en la provincia de Tucumán (San Miguel de Tucumán).

Introducción

El data mining está compuesto por numerosas técnicas que tienen como objetivo el buscar patrones e información relevante y oculta dentro de los datos almacenados en grandes bases de datos [6]. Actualmente con los avances en tecnología se han desarrollado variadas aplicaciones que permiten y facilitan la aplicación de técnicas de data mining y el procesamiento de grandes volúmenes de datos que son la materia prima para esta disciplina. La aplicación de estas técnicas como así también la utilización de las herramientas disponibles requiere que el usuario cuente con un cierto grado de conocimiento sobre el tema para la correcta selección y aplicación de las mismas. En los últimos años se ha venido desarrollando el área de aplicación de estas técnicas en datos obtenidos de procesos educativos con la utilización de diferentes medios tecnológicos, entre los que se encuentra la mediación con internet como una forma de facilitar el acceso y aumentar la carga horaria de las asignaturas. Los numerosos accesos para la realización de actividades, las relaciones que se crean y desarrollan, como las interacciones entre alumnos y docentes generan este gran volumen de datos de los cuales pueden extraerse información que correctamente interpretada puede favorecer al mejoramiento del proceso educativo guiando al docente en la toma de decisiones. Esta rama del data mining que procesa datos educativos se denomina Educational data mining (EDM) [1], y tiene como objetivo el descubrimiento de información a través de la aplicación de las técnicas a los datos relacionados con el desempeño pedagógico generados en plataformas LMS y también los datos de las actividades presenciales realizadas por los alumnos. Se desea analizar estos datos para extraer información por diversas razones como lo son:

- comprender la forma en que aprenden los alumnos.
- descubrir la mejor forma de organizar los materiales y actividades de la asignatura,
- determinar cuáles atributos son más representativos para ser utilizados en las predicciones.
- desarrollar nuevas tipologías de estudiantes y ajustar las existentes.
- determinar la formación de grupos.
- descubrir patrones de comportamientos.
- modificar las estrategias pedagógicas.

entre varias otras.

La aplicación de las técnicas de data mining permite superar las dificultades que presentan otras metodologías como las estadísticas en cuanto al manejo de grandes volúmenes de datos como así también al elevado número de variables que deben analizase durante el procesamiento. El análisis de datos educativos puede realizarse desde distintos enfoques de acuerdo a las técnicas que se apliquen, estas son:

Descripción: Este grupo presenta como principal objetivo el de describir las características más representativas de los datos en busca de un modelo que englobe los diferentes tipos analizados. Cuando estas técnicas se aplican a datos educativos lo que hace es caracterizar a los alumnos según su desempeño académico en las actividades propuestas por el docente relacionadas con el contenido específico de la asignatura [3].

Predicción: Las técnicas que se agrupan es este tipo buscan establecer un modelo que represente a los datos y que permita estimar los valores que pueden tomar a futuro las variables analizadas, en el caso de su aplicación a datos educativos los atributos iniciales de los alumnos son analizados por estas técnicas obteniéndose resultados que le predicen al docente como puede llegar a ser el desempeño pedagógico durante las actividades a realizar. Estas predicciones se basan en el análisis de los datos iniciales y resultados intermedios que van obteniendo los alumnos durante el desarrollo del contenido de la asignatura [4].

Segmentación: Se basan en encontrar conjuntos de datos con características

similares, conformando grupos de comportamientos similares y que sean diferentes entre sí. Cuando se procesan datos educativos los grupos se encuentran separados basándose en el comportamiento de los alumnos que presentan durante el desarrollo de las actividades de la asignatura [5].

Cada grupo de técnicas se aplica y representa un método o un enfoque conceptual para extraer la información de los datos. Estas pueden implementarse por medio de varios algoritmos que indican los pasos a seguir para la aplicación de cada técnica. Gracias a ellas pueden predecirse resultados generando modelos predictivos o encontrar relaciones que generan modelos descriptivos. Los resultados del procesamiento de los datos educativos con data mining tienen como destinatario al docente, que además, desde el punto de vista de la conveniencia representa el mejor usuario para realizar el procesamiento por participar del proceso y ser quien conoce en profundidad el campo de la disciplina. Como podemos observar tanto las técnicas como su aplicación tienen en general un grado de dificultad que requiere conocimientos por lo menos básicos sobre data mining, por ello en ayuda al docente, se ve la necesidad de desarrollar un framework que oriente y lo ayude en esta tarea. El desarrollo de dicho framework permitirá guiar a los docentes interesados en aplicar las técnicas de data mining durante el procesamiento de los datos, sugiriendo las técnicas más adecuadas, estableciendo prácticas y criterios para la interpretación de los resultados. Se espera que el aporte del data mining educativo al proceso, mejore en gran medida su efectividad. Permitiendo al docente autoregular los contenidos, las practicas pedagógicas como así también poder acercarle herramientas que permitan una clasificación a priori del perfil dominante de los alumnos como así también presentar los patrones emergentes de las actividades desarrolladas durante el cursado de la asignatura.

Metodología

Se pretende con esta investigación el generar un marco de referencia para la consulta de docentes interesados en innovar en el proceso educativo. El estudio de los resultados que presentan de la aplicación de las diferentes técnicas aplicándolas a datos educativos serán valorados desde el punto de vista del docente. Particularmente para el estudio se utilizarán datos de asignaturas con contenidos de programación de las diferentes carreras y subsedes de la facultad de Ciencias Exactas y Naturales de la Universidad Nacional de Catamarca. Como finalización de esta investigación se estima el poder desarrollar un framework que pueda ser aplicado por docentes para su validación.

Conclusión

Se espera establecer un conjunto de conceptos, prácticas y criterios que permitan procesar e interpretar la información obtenida de la aplicación de técnicas de data mining a datos obtenidos del proceso educativo en asignaturas relacionadas con contenidos de programación. Con el objetivo principal mejorar el proceso educativo mediante una adecuada modificación de metodologías y técnicas relacionadas con las características específicas del perfil dominante de los alumnos cursantes actuales.

Líneas de investigación y desarrollo

El presente trabajo se enmarca dentro de la investigación realizada para el desarrollo de la tesis de Maestría en Ciencias de la Computación, en donde se investigan los aportes del data mining al proceso educativo. Con el objetivo de mejorar dicho proceso, se presentan nuevos enfoques para el procesamiento de datos educativos utilizando técnicas y herramientas de data mining. Particularmente se trabaja con datos obtenidos del desarrollo de las actividades de asignaturas con contenido de programación por lo que presentan características distintivas de datos de otras asignaturas de la misma carrera.
El tema que se presenta viene profundizándose mediante el estudio continuo y con la

presentación de diferentes trabajos en reuniones científicas en donde se muestran los avances realizados y los posibles resultados que se esperan de la investigación. La propuesta se enmarca dentro de los temas de investigación del Centro de Estudios Estadísticos creado en la Universidad Nacional de Catamarca, Facultad de Ciencias Exactas y Naturales (Catamarca, Argentina).

Resultados y Objetivos

Los resultado esperados para esta investigación son los de establecer un conjunto de de conceptos, prácticas y criterios para el procesamiento e interpretación de los resultados del procesamiento de datos educativos con técnicas y herramientas de data mining.

Tiene como objetivos principales:

- Conformar un conjunto de técnicas de data mining que permitan mediante su aplicación la obtención de información para la toma de decisiones por parte del docente.
- Procesar los datos educativos con la finalidad de comprender la forma en que aprenden y realizan las actividades los alumnos.
- Obtener sugerencias para mejorar el proceso educativo en relación con el aprendizaje de la programación.
- Establecer los criterios de selección de datos que aportan más información en el procesamiento para establecer el perfil dominante de los alumnos en cada cohorte.
- Establecer criterios para la comprensión de la información que proporcionan las técnicas de data mining por tratarse de un caso especial como lo son los datos educativos.
- Determinar el método de selección de atributos que sean más representativos de las características de los alumnos que cursan asignaturas con contenidos de programación.

Formación de Recursos Humanos

Los autores del trabajo se encuentran en la etapa de desarrollo de sus tesis de posgrado en carreras relacionadas con el tema de investigación, como la Maestría en Ciencias de la Computación en donde el Mgter. Marcelo Sosa desarrolla actualmente la tesis denominada: "***Aportes de data mining a la mejora del proceso educativo con blended learning: formalización y experimentaciones***" bajo la dirección del Dr. Carlos Chesñevar perteneciente a la Universidad Nacional del Sur (*U.N.S*). Además el investigador se encuentra en la etapa de planificación de su tesis doctoral en el área de minería de datos en el Doctorado en Ciencias dictado en la Facultad de Ciencias Exactas y Naturales (*Fa.C.E.N.*) en convenio con la Universidad Nacional del Sur (*U.N.S*).
La Docente Investigadora Lic. Eugenia Sosa Bruchmann desarrolla su tesis en la carrera Especialización en Ingeniería en Software de la Universidad Nacional de San Luis denominada "***La experiencia del usuario desde un nuevo enfoque para el desarrollo de productos interactivos: el comportamiento emocional del usuario y la importancia de los atributos estéticos***" dirigida por el Dr. Germán Montejano. Los docentes investigadores desarrollan actividades de dirección de tesis de la carrera de Licenciatura en Tecnología Educativa de los siguientes alumnos: Varela Marino del Valle cuya tesis se denomina ***"Análisis del impacto de un aula virtual en el proceso de enseñanza y aprendizaje en la Escuela de Educación Técnica Nº 7 " José Alsina Alcobert",*** y Martín Fabián Molina cuya tesis se denomina: ***"Estudio de la implementación del uso de TIC´s en la enseñanza de la educación vial en el sistema educativo municipal de San Fernando del Valle de Catamarca".***

Además desarrollan las siguientes actividades:

- Dirección de proyectos de investigación de voluntariado y pertenecientes a la facultad a la cual pertenecen.

- Integración de equipo de investigación de centro de investigación en Estadística de la Facultad de Ciencias Exactas y Naturales del U.N.Ca.
- Producción de artículos científicos para su presentación en congresos locales, nacionales e internacionales.
- Participación de los integrantes en cursos de actualización y posgrado en el área de estudio.
- Integrantes de la revista de ciencias de la Facultad de Ciencias Agrarias de la U.N.Ca.
- La actualización y capacitación permanente de los investigadores en talleres o workshop relacionadas con el tema del trabajo.
- La participación de los investigadores como consultores en proyecto afines que se desarrollan en la Facultad de Ciencias exactas y Naturales en distintas áreas.
- Examinadores de trabajos finales en las diferentes carreras que se dictan en la Fa.C.E.N. de las U.N.Ca.
- Dirección de tesis y tesinas finales de las carreras de computación, informática y Licenciatura en tecnología educativa.
- La planificación de seminarios para docentes en temas relacionados con la investigación y resultados obtenidos en la investigación.
- Participación en convenios con la Facultad de Tecnología para el desarrollo de estudios del área de datamining.

REFERENCIAS

1. *C. Romero and S. Ventura, "Educational data mining: A Survey From 1995 to 2005", Expert System with Applications, vol. 33, pp. 135-146, 2007.*
2. *Lavrac, N., Kavsec, B., Flach, P todorovski, L., "Subgroup discovery with CN2-SD". Jornal of machine learning research. 2004.*
3. *Jain A.K.and Dubes R.C."Algorithsm for clustering data. 1998. Englewood Cliffs. N.J. Prentice Hall.*
4. *Fayyad, U, Piatetsky-Shapiro, G. and Smyth, P., "The KDD process for extracting usefull knowledge from volumes of data". Communication of ADM 1996.*
5. *Agrawal, R., Imiellinsky, T. and Swami, A.N.."Mining Assosiation Rules between set of item in large databases". In International conference on management of data. 1993. Whashington D.C. ACM Press.*
6. *Solarte Martinez, Guillermo Roberto, Ocampo S., Carlos Alberto. Técnicas de clasificación y análisis de representacion del conocimiento para problemas de diagnósticos. Cientia Et Technica [En Linea] 2009, Xv (Agosto-Sin Mes) : [Fecha De Consulta: 2 de marzo de 2016].*
7. *Gabriel Páramo and Carlos Correa, "Deserción Estudiantil Universitaria. Conceptualización," Medellín, Revista Abril - Mayo – Junio 1999.*
8. *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, pp. 37-54, 1991.*
9. *Daniel T Larose, Discovering Knowledge in Data: An Introduction to Data Mining. New York: John. Wiley & Sons, 2004.*
10. *Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students' Performance," in International Journal of Advanced Computer Science and Applications, India, 2011, pp. Vol. 2, No. 6.*
11. *Baker Ryan and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," JEDM - Journal of Educational Data Mining, vol. 1, no. 1, Octubre 2009.*
12. *Jonathan E. Freyberger, Neil Heernan, and Carolina Ruiz. Using association rules to guide a search for best fitting transfer models of student learning. Master's thesis, Worcester Polytechnic Institute, 2004.*
13. *Merceron and K. Yacef. Educational data mining: a case study. Process of 12th. Conference on Artificial Intelligence in Education (AIED03), page 467 .2005.*
14. *María Delia Grossi."Reglas de predicción aplicables al diseño de un curso de computación". Marzo 2008.*
15. *Erwin Sergio Fischer Angulo."Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios". Santiago de Chile 2012.*
16. *Pedro Gonzalez Garcia,."Aprendizaje evolutivo de reglas difusas para la descripción de subgrupos". Granada España. 2007.*
17. *Cristoban Romero, Sebastian Ventura, Nykola Pechenizkiy and Rayan Beker."Handbook of educational data mining".Chapman & Hall CRC press. 2011*